CGF Value-Added Data Analysis
Data Input Requirements
08/30/06

The CGF has recently updated our analytic capacity to provide investigators the option to receive more useful analyses upon delivery of their genotype data.

These value-added reports and analyses include:

- Completion rate per sample and assay
- Concordance among duplicates[1]
- Discordances between reported and genotyped sex
- Tests for deviation from Hardy-Weinberg proportions
- Allele and genotype frequency distributions
- Basic contingency table association analyses

In order to receive the results of these analyses, investigators must provide a data file listing duplicate and simple phenotypic information (i.e. case/control status and ethnicity). Otherwise, analysis reports will only contain completion rates per sample and assay.

Please note that this information is required only if an investigator or research group wishes to receive the most informative reports and analyses upon delivery of their genotype data.

Data must be provided in an Excel file with a single worksheet of data, or a Comma Separated Value (CSV) text file. The first row contains column headings and the subsequent rows contain data. The columns are as follows:

Column	Heading	Description
1	SID	Investigator specified unique sample identifier. In most cases, this is the NCI vial label provided in the sample manifest when samples are received at CGF.
2	PID	Investigator specified individual identifier. This information is used only to determine groups of samples that should correspond to the same individual. All samples with the same PID are assumed to be duplicates and will be included in the duplicate checking procedure[1]. Fictitious or anonymized identifiers are suggested, though the CGF will never reveal or share these identifiers.
3	PARENT1	Parent 1 and 2 individual identifiers. If the individual is a founder, then both PARENT1 and PARENT2 must be blank. Otherwise both PARENT1 and PARENT2 must contain a PID. When specifying complex pedigrees, individuals that do not contribute samples may be included as connectors by leaving the SID blank.
4	PARENT2	
5	SEX	Sex of the individual. Values must be "MALE", "FEMALE", or blank for unknown or not otherwise specified.
6	POPGROUP	Population group. This field specifies the population group that the sample belongs to. The values are user defined strings. E.g., "CAUCASIAN", "AFRICAN AMERICAN", "CENTER1", "CENTER2", etc. These groups will be used to perform simple stratified tests of Hardy-Weinberg proportions.
7	PHENOGROUP	Phenotype group. This field specifies categorical phenotypes. The values are user defined strings. E.g., "CASE", "CONTROL", "EARLY", "ADVANCED", "QC", etc. These groups will be used to perform simple stratified association tests.

The file should be sent as an email attachment to Meredith Yeager (yeagerm@mail.nih.gov) and Jeff Yuenger (yuengerj@mail.nih.gov).

Example data file

SID	PID	PARENT1	PARENT2	SEX	POPGROUP	PHENOGROUP
AG 512 123	I001			MALE		QC
AG 512 124	I001			MALE		QC
AG 513 123	I002			FEMALE		QC
AG 513 124	I002			FEMALE		QC
AG 514 123	I003	I001	I002	MALE		QC
AG 514 124	I003	I001	I002	MALE		QC
AG 515 123	I004			MALE	CAUCASIAN	CASE
AG 515 124	I004			MALE	CAUCASIAN	CASE
AG 516 123	I005			MALE	CAUCASIAN	CONTROL
AG 517 123	I006			MALE	CAUCASIAN	CONTROL
AG 518 123	I007			MALE	CAUCASIAN	CASE

More than 7 columns can be accommodated if necessary (e.g. POPGROUP1, POPGROUP2, etc). For POPGROUP and PHENOGROUP columns, please use descriptive names and/or provide a key to define the text fields in a separate file (e.g. “1 = case, 0 = control”).

[1]Identifiler data may also be used to identify duplicates and validate pedigree relationships.